Understanding facial impressions between and within identities

Mila Mileva, Andrew W. Young, Robin S. S. Kramer[1] & A. Mike Burton
Department of Psychology, University of York, UK

Correspondence to:
Mila Mileva
Department of Psychology
University of York
Heslington, York
YO10 5DD, UK
mila.mileva@york.ac.uk

Running head: Facial impressions

Keywords: Social perception; facial impressions; face perception; trustworthiness; attractiveness; dominance

Funding: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.323262 to Mike Burton.

[1] Present address: School of Psychology, University of Lincoln.

Abstract

A paradoxical finding from recent studies of face perception is that observers are error-prone and inconsistent when judging the identity of unfamiliar faces, but nevertheless reasonably consistent when judging traits. Our aim is to understand this difference. Using everyday ambient images of faces, we show that visual image statistics can predict observers' consensual impressions of trustworthiness, attractiveness and dominance, which represent key dimensions of evaluation in leading theoretical accounts of trait judgement. In Study 1, image statistics derived from ambient images of multiple face identities were able to account for 51% of the variance in consensual impressions of entirely novel ambient images. Shape properties were more effective predictors than surface properties, but a combination of both achieved the best results. In Study 2 and Study 3, statistics derived from multiple images of a particular face achieved the best generalisation to new images of that face, but there was nonetheless significant generalisation between images of the faces of different individuals. Hence, whereas idiosyncratic variability across different images of the same face is sufficient to cause substantial problems in judging the identities of unfamiliar faces, there are consistencies between faces which are sufficient to support (to some extent) consensual trait judgements. Furthermore, much of this consistency can be captured in simple operational models based on image statistics.

Introduction

The faces we encounter in our everyday lives can be very variable in appearance, depending on pose, expression, illumination and other factors. For example, Figure 1 shows different images of the same person. Jenkins, White, Van Montfort and Burton (2011) used the concept of 'ambient images' to refer to these types of image and the great amount of variability they reflect. Our interest here is in the consequences of this variability, and especially its implications for the perception of identity and for social trait impressions.

Figure 1: *Example ambient images of the same face from one of the identities used in Studies 1-3. The depicted identity has given permission for his images to be reproduced here.*

Implications for the perception of face identity are beginning to be relatively well-understood. For most perceivers, variability in the images falling on the retina creates remarkably little difficulty for recognising the identities of highly familiar faces (Bruce & Young, 1986; Burton, Jenkins, Hancock & White, 2005; Young & Burton, 2017). For faces of unfamiliar people, however, matters are very different. If you do not know the person shown in Figure 1, it can be tricky even to see that these are all photos of the same face (Jenkins et al., 2011). As a consequence, performance in matching and recognition tasks involving ambient images of unfamiliar faces is generally susceptible to substantial error rates (Hancock, Bruce & Burton, 2000; Young & Burton, 2018), though there is a

surprisingly wide range of performance across different observers in the normal population (Burton, White & McNeill, 2010).

By analysing the statistical properties of ambient images of faces, Burton, Kramer, Ritchie and Jenkins (2016) showed that image variability is to some extent idiosyncratic - that is, the ways in which one person's face varies across images can be different for someone else's face. Learning to recognise a familiar face thus involves learning how that face can vary, through seeing it in many settings – in effect becoming sufficiently expert with that face identity to be able to recognise new photos of the same person. But this form of perceptual expertise is identity-specific and may not generalise to another person's face, because that face varies in different ways. For this reason, unfamiliar face recognition is often poor because the range of variability of an unfamiliar face is unknown (Burton et al., 2016; Kramer, Young & Burton, 2018; Young & Burton, 2018). Computational work has shown the utility of this approach by simulating a range of well-known properties of familiar and unfamiliar face recognition (Kramer, Young, Day & Burton, 2017; Kramer et al., 2018).

Although perception of the identities of unfamiliar faces can be problematic, many other characteristics are more easily seen. These include relatively objective properties determined by structural cues such as apparent gender, age and ethnicity (Bruce & Young, 2012) and more subjective impressions of social dispositions such as friendliness or trustworthiness. These impressions based on facial appearance are what we particularly seek to understand here. They influence actions ranging from whether to approach someone at a party to whether to vote for them in an election or judge them guilty of a crime (Olivola, Funk, & Todorov, 2014; Todorov, Olivola, Dotsch & Mende-Siedlecki, 2015). Although they are known to be of limited validity (Todorov, 2017), such impressions are to some extent consensual across different perceivers (Kramer, Mileva & Ritchie, 2018; Oosterhof & Todorov, 2008; Sprengelmeyer et al., 2016; Sutherland et al., 2013) and they can be formed from little more than a single glance (South Palomares & Young, 2018; Willis & Todorov, 2006). Understanding how they are created has been a focus of much recent interest.

An important advance has been to demonstrate that, while they can involve many different traits, facial impressions mainly fall along a relatively small number of underlying evaluative dimensions. Oosterhof and Todorov (2008) found dimensions that approximated perceived trustworthiness and dominance, and later studies have both replicated this underlying pattern

and suggested that youthful attractiveness may form a third dimension (South Palomares, Sutherland & Young, 2018; Sutherland et al, 2013). These accounts explain how a wide variety of impressions can be derived from a relatively simple underlying structure, since perceived traits and attributes can be evaluated from their positioning in the resulting three-dimensional space (Oldmeadow, Sutherland & Young, 2013; Oosterhof & Todorov, 2008; Sutherland et al., 2013).

Despite this relatively simple structure to trait impressions, it is clear that multiple covarying visual cues are used, with no single cue completely controlling a given dimension (Santos & Young, 2011; Todorov, 2017; Young, 2018). For example, smiling can make a face look relatively trustworthy, attractive, or even in some circumstances dominant, depending on the type of smile and the way it is combined with other cues (Young, 2018). These interacting cues can be visualised and manipulated with data-driven techniques involving computer modelling (Oosterhof & Todorov, 2008; Walker & Vetter, 2009) or computer image manipulation (Sutherland et al., 2013; Sutherland, Rhodes & Young, 2017). Interestingly, many of the cues are highly image-dependent, such that across different ambient images the same face can look trustworthy or untrustworthy, attractive or unattractive, dominant or submissive, depending on pose, expression, lighting and other variables (Jenkins et al., 2011; Sutherland, Young & Rhodes, 2017; Todorov & Porter, 2014). In sum, impressions of the same face can vary considerably - a person can look approachable at one moment and forbidding the next moment, as is evident in Figure 1. However, perceivers will often misattribute these momentary dispositions as reflecting stable traits of unfamiliar people (Todorov, 2017).

There is a striking difference, then, between the way that image variability needs to be used in determining face identity and in forming trait impressions. Identity needs to be recognised despite differences between images of the same face, whereas trait impressions need to make use of these image differences and in consequence are highly image-dependent. Moreover, and seemingly paradoxically, image differences are interpreted inconsistently by different observers when used to perceive unfamiliar face identity, yet the same image differences are interpreted relatively consistently by different observers when forming trait impressions.

Of course, judgements of identity and judgements of traits also have very different functional roles in our daily lives that will shape the functional organisation of an optimal face

perception system (Young, 2018). Identity is fundamentally driven by the requirement to recognise people we know, whereas trait judgements based on faces are, by definition, particularly useful for unfamiliar people. For people we know, we can infer character from past behaviour rather than relying only on facial appearance. We might therefore expect that the physical cues signalling identity will, to some extent, be different from those signalling personal traits. It follows that, unlike face recognition, the cues characterising trait impressions are likely to be relatively consistent across many different faces (Young & Burton, 2017; Young, 2018).

Cue consistency has already been demonstrated for the perception of the relatively objective characteristic of face gender, where it has been established that a dimension that is a by-product of encoding the identities of familiar faces will also serve to classify any face image (including images of unfamiliar faces) as male or female (Kramer, Young et al., 2017). However, little is known about the possibilities of consistent versus idiosyncratic variability in the cues that underlie more nuanced and apparently subjective impressions. On the one hand, given the fact that observers form reasonably consensual impressions of images of unfamiliar faces, it seems likely that they use cues that can generalise across many face identities. On the other hand, the evidence of identity-specific image variability is compelling and there is evidence that an individual may (for example) have at least partially idiosyncratic facial expressions (Cohn, Schmidt, Gross & Ekman, 2002; Kaufmann & Schweinberger, 2004; Mileva & Burton, 2018), hence being able to interpret such identity-specific idiosyncrasies could confer an advantage in forming impressions of that person from their face.

Our first step in the present study was therefore to determine whether consensual impressions of everyday face images can be modelled directly from physical image properties (Study 1). To achieve this we used Principal Components Analysis (PCA) of a set of highly varied ambient images to extract PCs describing the underlying variation in shape and surface texture properties across the image set. In line with other image-based approaches (Burton, Miller, Bruce, Hancock & Henderson, 2001; Calder, Burton, Miller, Young & Akamatsu, 2001; Craw, 1995; Kramer, Jenkins & Burton, 2017), we defined image shapes in terms of the locations of fiducial positions marking the locations of facial features (eyes, nose, mouth etc). Surface texture properties (pixel colour and brightness values) were then derived across shape-normalised images, i.e. images that were reshaped in a manner that put each image's

fiducials into the same set of locations, so that the positions and shapes of features themselves were held constant (Burton et al., 2001; Calder et al., 2001; Craw, 2005). We then created regression equations that used these shape and surface texture PCs to model the rated trustworthiness, attractiveness and dominance of each image in a training set. As a strong test of its validity, we cross-validated each regression model by establishing how well it could predict the rated trustworthiness, attractiveness and dominance of a novel set of ambient images that had not been included in the training set of images used to create that model. This generalisation test is essential because it shows how well each model has captured genuinely informative rather than spurious covariation in the trained images.

Importantly, PCA is used here mainly as a convenient description of the statistical properties of a set of face images. We make no assumptions about whether the human visual system uses PCA - its value is instead that it offers a principled data reduction format and, of course, that the same technique was used by Burton et al. (2016) to investigate identity-specific variability.

Our second step was to evaluate the potential contribution of face identity to impression formation by incorporating within-person variability into our predictive models (Study 2 and Study 3). As noted above, studies of variability across different images of the same face have shown that this variability is to some extent idiosyncratic - the ways in which one person's face varies in appearance are different from how someone else's face will vary (Burton et al., 2016). Being able to represent the statistics of this idiosyncratic variability can assist in recognising the face of a familiar individual (Kramer, Young et al., 2017; Kramer et al., 2018), whereas lack of knowledge of the variability of unfamiliar faces hinders their recognition (Kramer, Young et al., 2017; Kramer et al., 2018; Young & Burton, 2018b). We therefore sought to determine whether being able to represent the identity-specific statistical variability of an individual face would assist the formation of impressions of social traits such as trustworthiness, or whether interpreting these traits could be learnt from any other face. So, while previous work has established the value of PCA to model first impressions (Oosterhof & Todorov, 2008; Walker & Vetter, 2009), here we add the novel step of establishing the statistics of multiple images of the same person. This way, we can sample both within- and between-person variability.

To achieve this, we created regression-based models from sets of multiple images of the same

face or from images of several different faces and tested whether within-identity models created from images of the same face were better able to predict the trustworthiness, attractiveness and dominance of novel images of that face than were cross-identity models created from different faces. In Study 2, the within-identity models were based on a single face, and in Study 3, on a small set of faces.

Study 1

In Study 1, we evaluated how much of the variance in impressions of trustworthiness, attractiveness and dominance of images of faces can be modelled from physical image properties. The face images we used were all unstandardised, everyday images of the type Jenkins et al. (2011) term 'ambient images'. As such, they are often thought to present a significant challenge to modelling because of the great many ways in which such images can vary (see Figure 1).

Method

Stimuli

The image set for Study 1 consisted of 20 images of each of 20 unfamiliar people (10 men), 400 in all. These were foreign celebrities and a relative of one of the authors, which ensured the availability of many images for each identity. All were unfamiliar to our participants. Foreign celebrity images were downloaded from an internet search by entering the name of the person and choosing images that were in full colour, with all facial features needed to position fiducials visible in the image, and with no parts of the face obscured by clothing or glasses. These were all naturally occurring ambient images (Jenkins et al., 2011) and included a large amount of variability due to lighting, pose and expression for each identity (see Figure 1 for examples).

Participants

Images were rated by 20 participants (mean age = 20.1 years, age range = 18-24 years), all from the University of York. Sample size was based on Todorov and Porter (2014) who also collected ratings from 20 participants per image. All participants had normal or corrected-to-normal vision and received payment or course credit for their participation. Participants

provided informed consent in accordance with the ethical standards of the 1964 Declaration of Helsinki. Experimental procedures were also approved by the Ethics Committee of the University of York Psychology Department.

Image rating task

The rating task was computer-based, and stimuli were displayed on an 18-inch LCD monitor. The experimental program was written in MATLAB and used functions of the Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Face images were presented individually at the centre of the screen with a rating scale positioned underneath. Participants were asked to rate each image on a scale from 1 (not at all trustworthy/attractive/dominant) to 9 (extremely trustworthy/attractive/dominant) using a mouse-click. The task was self-paced, with an inter-stimulus interval of 1s. As is common in first impressions research, participants were not given detailed instructions as to how to interpret the rating dimensions (trustworthiness, attractiveness, dominance), but were instead encouraged to rely on their "gut instinct" (cf. Todorov, Mandisodza, Goren, & Hall, 2005). They were also informed that they would see multiple different images of the same faces. Participants provided ratings for all 400 images. Each image was rated for a single social attribute (trustworthiness, attractiveness, or dominance) across a block of 400 trials, and the order of the blocks corresponding to each of the three ratings was randomised to reduce any carryover effects (Rhodes, 2006). Order of image presentation was also randomised individually for each participant.

From these ratings of the trustworthiness, attractiveness and dominance of each image we derived the mean rating across all 20 participants for each of the 400 images for each of the 3 traits. These averaged ratings of each image formed the data to be modelled.

Image PCA and Regression Models

The procedure for PCA of the images followed that used by Burton et al. (2016), where full details can be found and appropriate software has been provided in Kramer, Jenkins and Burton (2017). Prior to PCA, images were scaled to 190 x 285 pixels and represented in RGB colour space using a lossless image format (bitmap). Face image shape was derived by manually aligning the points of a standard grid of 82 fiducial positions with anatomical landmarks. The positioning of this grid of fiducials on each image led to 82 xy-coordinates,

creating a shape-vector of 164 numbers (82 points x 2 coordinates) for each image. In order to derive image surface texture-vectors, the average location of each fiducial across the whole image set was calculated. The surface texture (RGB values) for each image was then morphed to this average shape, so that corresponding fiducials were aligned across different images and in consequence all 2D shape information resulting from differences between images in the fiducial locations themselves was discarded. This generated a texture-vector of pixel intensities comprising 162,450 numbers (190 width x 285 height x 3 RGB layers) describing these nominally 'shape-free' images. PCA was performed separately for shape (using the original fiducial locations) and for surface texture (based on the shape-free images). This generated a number of shape and texture eigenvectors (equivalent to Principal Components, also referred to as 'eigenfaces' in the literature).

Principal Components (PCs) are ordered by the amount of variation they account for, so early components explain more variability than later ones. Each PC captures a set of properties that describes a way in which different images vary and each face image can be represented as a linear combination of these eigenfaces, providing each image with a unique set of coefficients (or weights) that acts as its signature. For the purposes of the present study, the first 30 PCs for shape and for surface texture were used to model the variability of a set of training images, with each individual image coded as a set of 30 shape and 30 texture coefficients.

In order to test the validity of the regression-based models we developed, PCA was applied separately across 5 different subsets of 320 images from the full set of 400. Each subset of 320 model training images used 16 of the 20 images for each of the 20 faces comprising the full set, allowing 80 images (4 images for each of the 20 people) to be held back to form an independent cross-validation test of the generalisability of the regression model to new ambient images. In order to ensure a fair comparison of levels of performance between trained images and novel test images, we measured each model's performance across the 320 trained images using a random sample of 80 of these, thus equating the number of images used to measure performance in each case (i.e. 80 images from the training set and 80 untrained novel test images). This procedure was repeated 5 times, each involving a different sample of trained and untrained test images.

The regression models were based on stepwise linear regression with the averaged ratings of a given trait (trustworthiness, attractiveness, or dominance) for each image as the dependent

variable and image shape and surface texture coefficients as predictor variables. A separate regression was used for each trait. We used these models to estimate the proportion of variance in the overall group ratings of trustworthiness, attractiveness or dominance for each image that could be accounted for in the trained images (images subjected to PCA) and then in the untrained test images used to assess the model's generalisability.

Results and Discussion

As is often the case in this type of research, our analyses are based on the mean rating per image across all participants, not on the individual participants' ratings themselves. That is, we are modelling the consensual (average) component of impressions of each image. As is also common in this type of research, inter-rater agreement was high for all three social attributes (Cronbach's alphas above .90). This measure is appropriate when drawing conclusions about the overall group values, because alpha measures the extent to which a group of items (here, observers) will agree with other groups, tested in future (Cortina, 1993). High values of alpha therefore show that ratings of each image are likely to be stable across groups of participants.

Ratings of attractiveness (D (400) = .08, $p < .001$) and dominance (D (400) = .06, $p = .005$) were not normally distributed, therefore we also calculated Kendall's W as a non-parametric alternative to alpha. Here the values were lower, indicating that there were differences between observers, but nonetheless revealed an underlying core of significant agreement (Attractiveness, $W = 0.50$, $p < .001$; Trustworthiness, $W = 0.21$, $p < .001$; Dominance, $W = 0.23$, $p < .001$). We return to the implications of this later.

Figure 2 shows mean ratings for all images on the three social judgements. These are displayed separately for male and female face identities, and ranked by overall mean, separately for each judgement. Again consistent with previous work (Jenkins et al, 2011; Sutherland, Young et al., 2017; Todorov & Porter, 2014), there are substantial differences in the ratings given to different images of the same person, and this is true for all three judgements. This is evident even after ordering each identity by their mean score, which would emphasise any between-person differences. That said, it is clear that differences in the average overall attractiveness across different faces were a little more pronounced than were differences in perceived trustworthiness or dominance and the same trend is also evident in other published work (Sutherland, Young et al., 2017; Todorov & Porter, 2014). In fact, it is

always possible to pick images of any two individuals where one would be perceived as more trustworthy or dominant than the other.

The data in Figure 2 are separated by face gender simply to show that the extent of variability is not gender-specific. Other than the slightly lower attractiveness ratings for some of the male faces there are no obvious gender differences, and face gender was ignored for the PCA.

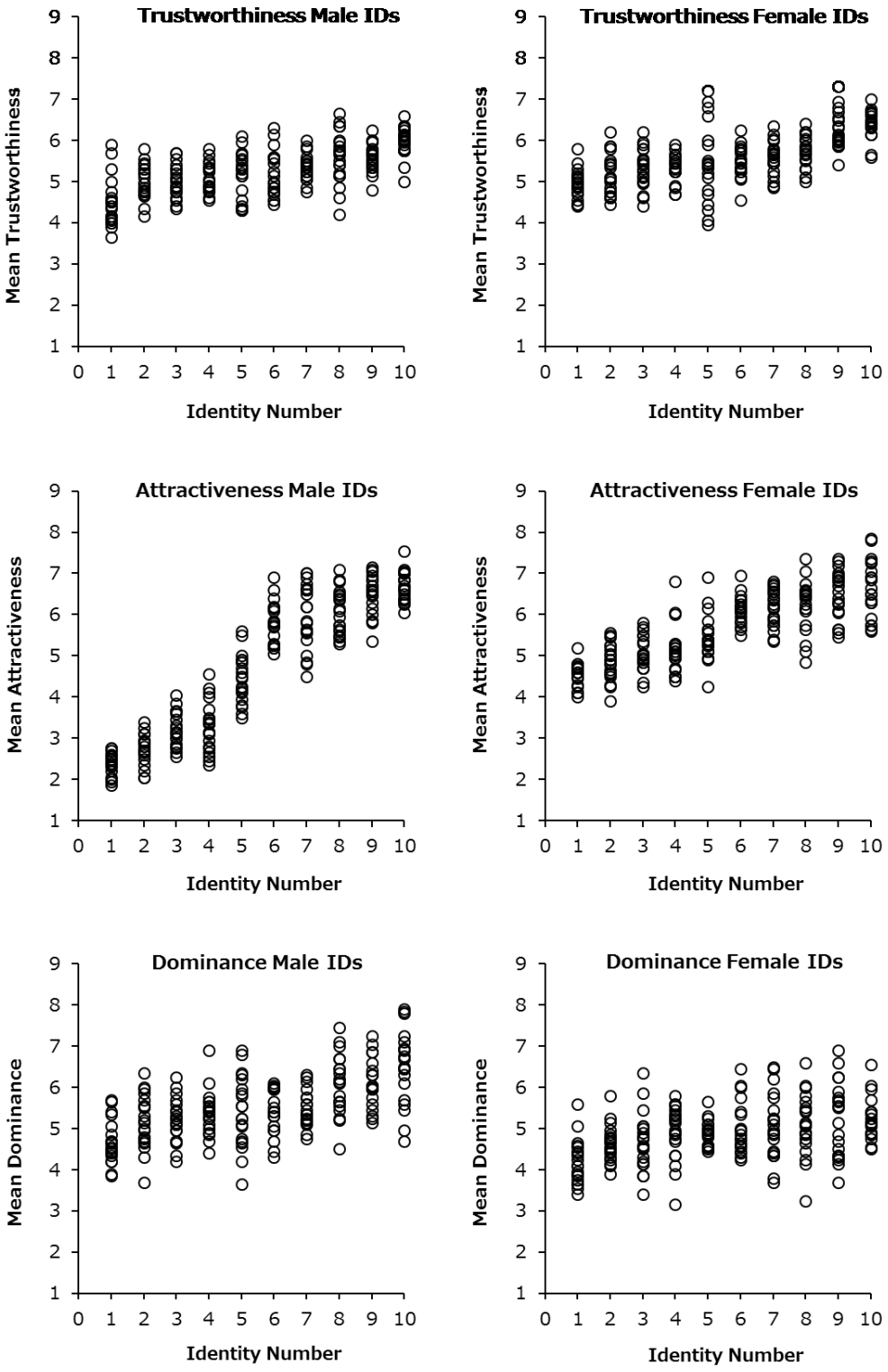

Figure 2: *Mean ratings of all images from the Study 1 set of 20 images of each of 20 face identities for trustworthiness (top), attractiveness (middle) and dominance (bottom), displayed separately for male (left) and female (right) face identities. Each column represents*

a single identity and each point represents a single image. Identities are ranked on the x-axis by mean identity score, separately for each rating. There are substantial differences in the ratings given to different images of the same person for all three judgements.

These observations of consistency in overall ratings of each image (high values of Cronbach's alpha and highly significant values of Kendall's *W*) and substantial variability across the items themselves (Figure 2) confirm that the selection of images used shows properties similar to those noted in previous studies using ambient images (Sutherland, Young et al., 2017; Todorov & Porter, 2014). Having established this, we turn to whether the physical properties of the images can be used to predict the overall impressions of trustworthiness, attractiveness and dominance.

Using image properties to predict impressions

We used up to 60 derived dimensions from PCA (30 shape, 30 surface texture) to model the average social trait ratings of trustworthiness, attractiveness and dominance of each image through stepwise linear regression. In each case, 320 images were used to create the regression model (from which 80 were randomly selected to estimate performance with trained images), whilst the remaining 80 images served as a novel test set to determine the model's generalisability. This procedure was repeated 5 times for each trait, using different sets of trained and novel test images. Regression models were created from the top 5, 10, 15, 20, 25 and 30 PCs representing fiducial shape and surface texture components, and for the corresponding combinations of shape and texture components (involving 10, 20, 30, 40, 50 or 60 PCs). The first 30 shape and the first 30 texture components explained 99.6% and 86.6% of the overall variance on average across the 5 iterations of the procedure.

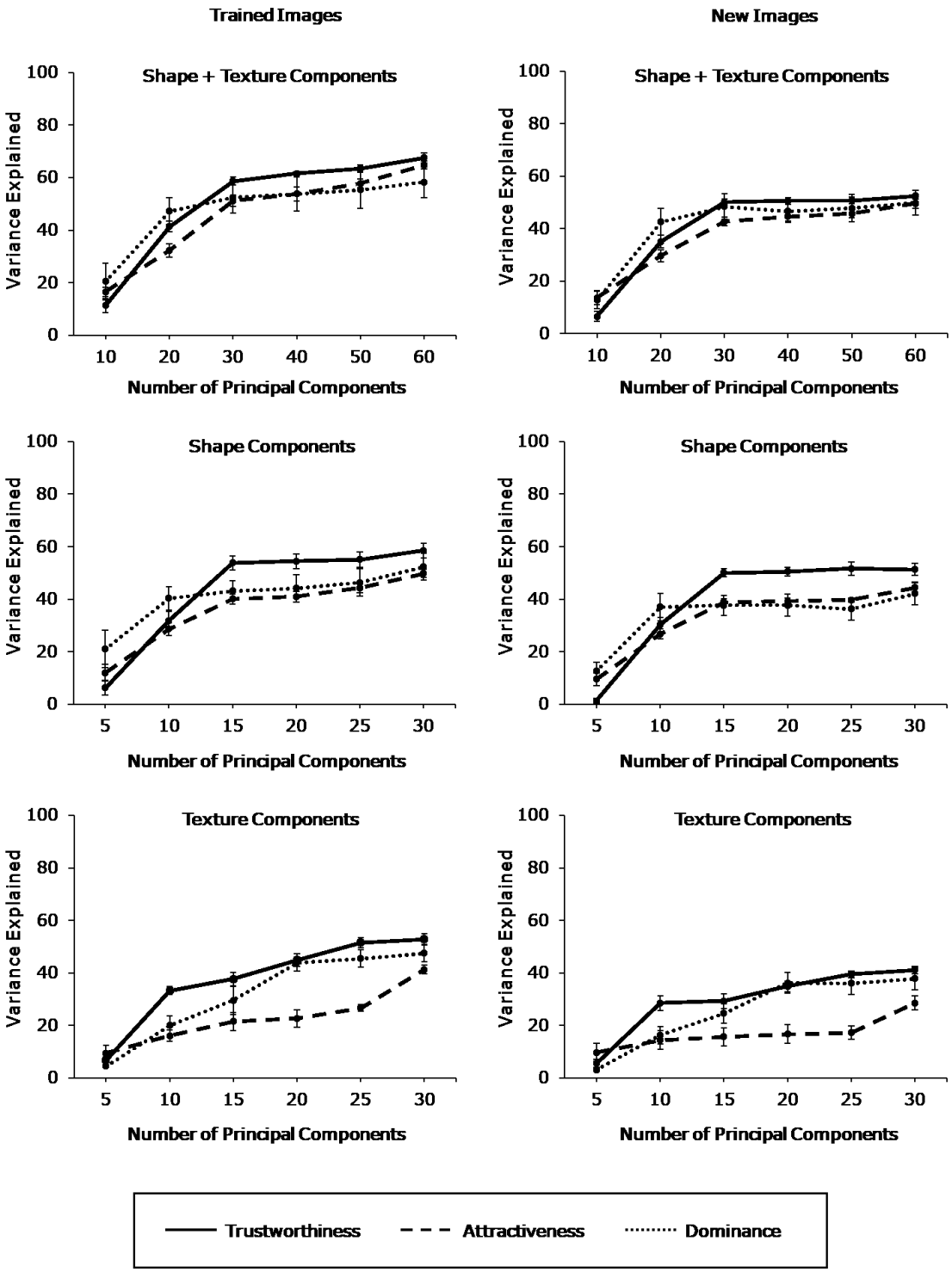

Figure 3: *Percentages of variance in participants' impressions of each attribute in Study 1 that can be explained using different numbers of shape and surface texture PCs for sets of trained ambient images used to create each regression model (left column) and in the cross-validation generalisation tests involving untrained novel images (right column). Mean*

performance across five separate iterations is shown, and error bars represent standard error. The upper row shows models that combine shape and surface texture PCs, the middle row uses only shape PCs, and the lower row uses only texture PCs.

Figure 3 shows the proportions of variance in participants' impressions of each attribute that can be explained from the physical properties of the images used to create each regression model and in the cross-validation generalisation tests involving novel images. It is clear that models based on an analysis of image statistics using a combination of shape and surface texture PCs are reasonably successful, being able to account for 64% of the variance in impressions across the modelled images and 51% across untrained novel images. Performance with the novel images does not improve much from using more than 30 PCs from the combined shape and texture predictor set.

Much of this success can be attributed to the shape PCs, which are themselves able to account for 54% of the variance in impressions across the modelled images and 46% across novel images. In contrast, surface texture PCs are less informative overall but still able to account for 47% of the variance in impressions across the modelled images and 36% across novel images.

To provide statistical confirmation of these main points we used Wilcoxon matched-pairs signed ranks tests to compare peak levels of performance. For the trained images, peak performance for models that combined shape and surface texture PCs was better than for shape PCs only ($Z = 3.41$, $p = .003$) or for surface texture PCs only ($Z = 3.41$, $p = .003$), and performance with shape PCs was also better than for surface texture PCs ($Z = 2.90$, $p = .004$). The same pattern was evident for performance with the novel test images; shape+texture > shape only, $Z = 2.44$, $p = .015$, shape+texture > texture only, $Z = 3.41$, $p = .003$, shape only > texture only $Z = 2.90$, $p = .008$.

A caveat concerning the above analyses might be that they are based on 60 PCs for shape+texture models and only 30 PCs for the shape only or surface texture only models. We therefore repeated the appropriate comparisons using performance levels for only 30 PCs in the shape+texture model, with the following results; for trained images - 30 shape+texture > 30 shape PCs, $Z = .17$, $p > .05$, 30 shape+texture > 30 texture PCs, $Z = 3.12$, $p = .006$, for novel images - 30 shape+texture > 30 shape PCs, $Z = .34$, $p > .05$, and 30 shape+texture > 30 texture PCs, $Z = 3.41$, $p = .003$. The Holm-Bonferroni correction (1979) was applied to

account for the multiple comparisons in both analyses. Supplementary Table 1 shows full statistics for each level of 'N PCs'. The relatively weak performance from surface texture PCs alone was therefore borne out by these complementary analyses. We note that in Figure 3 this was particularly noticeable for estimating perceived attractiveness.

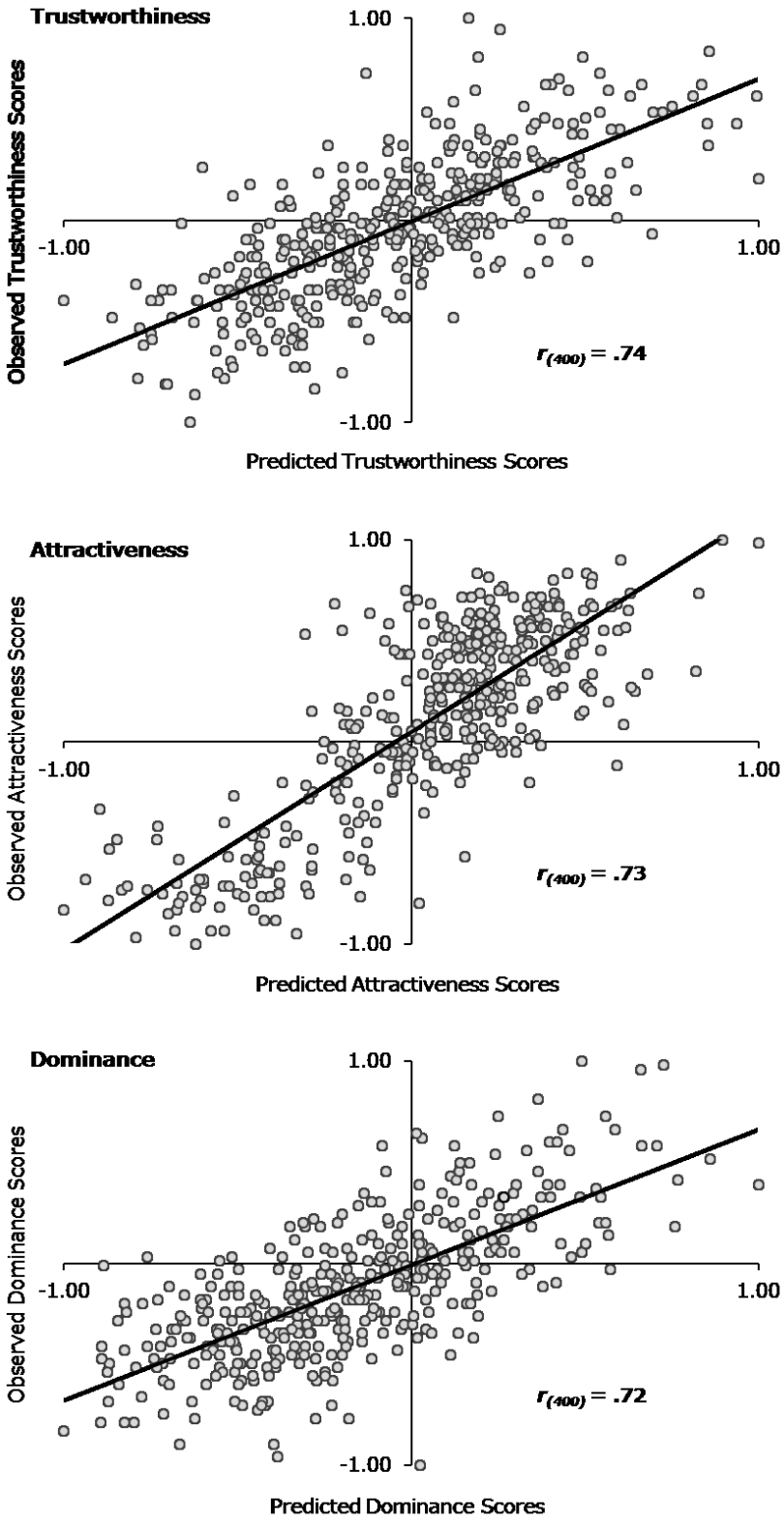

Figure 4: *Scatterplots showing the relation between observed ratings (averaged across human participants) and model-predicted ratings of the trustworthiness, attractiveness, and dominance of novel test images in Study 1.*

Our critical test of each regression model involved its ability to predict impressions of untrained ambient images. The design of Study 1 meant that across the 5 runs, all 400 images were used once each as untrained novel test items. Figure 4 shows scatterplots of the model-predicted ratings for these novel test items (using the combination of 60 shape and surface PCs that led to best generalisation performance) against the averaged observer ratings of trustworthiness, attractiveness and dominance. Substantial Spearman's correlations were found (trustworthiness, $r = 0.74$, $p < .001$; attractiveness, $r = 0.73$, $p < .001$; dominance, $r = 0.72$, $p < .001$), confirming that differences in impressions were being successfully modelled at the image level.

Study 2

Having shown in Study 1 that much of the variability in trait impressions can be modelled directly from image properties, in Study 2 we examined whether knowledge of how an individual face can vary is useful to forming trait impressions. This is an important issue because by applying PCA to multiple images of the same face, Burton et al. (2016) have demonstrated that the PCs that best describe the variability of one individual face are different from the PCs that best describe another face. That is, variability is to some extent identity-specific. This observation has been used to explain differences between familiar and unfamiliar face recognition, and in particular the way in which we become 'experts' at recognising individual familiar faces (Burton, 2013; Young & Burton, 2017, 2018a). Here we ask whether this identity-specific variability will also give rise to idiosyncratic social attributions, or whether the cues that form the source of these attributions are instead shared across different faces. To many researchers it seems intuitively likely that knowledge of an individual face may help to some extent in forming impressions. For example, a person may have a characteristic way of smiling that is more easily picked up in comparison with their other facial expressions (Cohn et al., 2002; Kaufmann & Schweinberger, 2004).

For Study 2 we tested the extreme cases in which a regression model was created entirely from ambient images of the same person's face or from a mixed set of different faces. To do this, we assembled larger samples of ambient images for each of a few faces, which could then be used to derive an identity-specific PCA of variability representing only one of these people. We could then use the regression technique with an individual face's PCA, to predict attributions made to new photos of that person. Performance with the same sets of novel

ambient images was directly compared across regression-based models created using the same face as the test images or created from images of a mixed set of different faces.

Method

Stimuli

We selected four identities (2 male) from the set used in Study 1, and collected 100 images of each. Selection criteria were the same as in Study 1: images were downloaded from an internet search on names, and the first 100 images returned (including those already used in Study 1) were chosen for which all facial features needed to position fiducials were visible and not obscured by clothing or glasses. We also used the 20 images of the 16 other faces included in Study 1 as additional stimuli to create sets of face images that included multiple identities.

Participants

Images were rated by 40 participants (mean age = 20.5 years, age range = 18-25 years), all from the University of York. All had normal or corrected-to-normal vision and received payment or course credit for their participation. Participants provided informed consent prior to their participation in accordance with the ethical standards of the 1964 Declaration of Helsinki. Experimental procedures were also approved by the Ethics Committee of the University of York Psychology Department.

Image rating task

Using the same general procedure as Study 1, all 400 of the newly collected images were rated for attractiveness, trustworthiness and dominance on a nine-point scale. Rating trials were arranged into counterbalanced blocks involving a single characteristic (trustworthiness, attractiveness, or dominance) and images were presented in a separate random order for each participant. Each participant rated 50 images per identity (200 in total) in such a fashion that all images were rated by 20 participants.

Image PCA and Regression Models

First, we conducted PCA on images of each target person separately. Each PCA was carried

out using the same general procedure as for Study 1, but the sets of ambient images subjected to PCA were now entirely of the same individual. Our intention was to establish, for each person, how well the variability in social impressions they create can be predicted from the image properties of their own photos. This approach differs from Study 1, in which predictive models were derived from image sets containing a mixture of both within-person and between-person variations (multiple images of multiple people). Instead, Study 2 included an exclusively 'within-identity' condition in which regression models were now created person by person and tested for generalisation to new images of the same face.

If we call the first face person A, we sampled 80 images of face A from the 100 available and ran the regression analysis as before, leaving aside the remaining 20 images of face A for testing the generalisation of the regression model. This procedure was repeated 5 times for face A, with each repetition involving a different set of 20 test items, thus leading to an estimate of performance following within-identity training across 5 face A test sets, measured with the 5 regression-based models.

As for Study 1, to ensure a fair comparison of levels of performance between trained images and novel test images, we measured each model's performance from a subset of the trained images; this time using a random sample of 20 of the 80 trained images to match the sample size for the untrained novel test images.

The same procedure was then followed for faces B, C, D. In this way, we modelled how much of the variance could be accounted for in perceptions of trustworthiness, attractiveness and dominance across 5 different runs involving trained and untrained images of the same individual, for each of 4 different faces.

To compare performance of these exclusively within-identity models to cross-identity models created from multiple faces of different individuals, we used 20 images of the 16 faces remaining from Study 1. This resulted in a total set of 320 varied-identity images that did not include faces A-D, which were then subdivided into four independent sets of 80 images (each with 5 images of 16 faces).

These sets of 80 images of multiple identities were then each subjected to separate PCAs and used to create regression models for the perception of trustworthiness, attractiveness and dominance in each set. We then tested how well these cross-identity models could predict

participants' impressions across 5 different sets of already trained images (i.e. images of the same person for the within-identity model and images of the 16 different identities used in the training for the cross-identity model) and, critically, across the 5 sets of untrained images for each of faces A, B, C, and D (where none of these face identities had been in the training set).

In this way, we were able to compare performance at predicting the attributes of untrained images (sets of 20) across regression models that were trained on sets of images of the test face (A, B, C, or D, comprising the within-identity condition) or on images of multiple other faces (a cross-identity condition). In other words, by comparing performance of within-identity and cross-identity models on the same sets of novel test images, we can directly measure whether training that encompasses the within-person variability of that particular face confers any benefit.

Results and Discussion

As in Study 1, high Cronbach's alphas were noted (above .88 for all three traits). Ratings of attractiveness (D (400) = .14, $p < .001$), trustworthiness (D (400) = .08, $p < .001$) and dominance (D (400) = .07, $p < .001$) did not follow the normal distribution, therefore we calculated Kendall's W as an alternative agreement measure. As the first 50 images of each identity were rated by one group of participants and the remaining 50 images of each identity by a different group, we calculated Kendall's W separately for each set. Similarly to Study 1, the values were lower than Cronbach's alpha but still showed significant agreement (Attractiveness, set 1 $W = 0.43$, $p < .001$, set 2 $W = 0.44$, $p < .001$; Trustworthiness, set 1 $W = 0.20$, $p < .001$, set 2 $W = 0.26$, $p < .001$; Dominance, set 1 $W = 0.11$, $p < .001$, set 2 $W = 0.14$, $p < .001$).

Figure 5 shows the spread of ratings for each trait across all four identities. Consistent with results from Study 1, there is a substantial range of variability in impressions of different images of each person, but with some overall between-person differences in ratings of attractiveness.

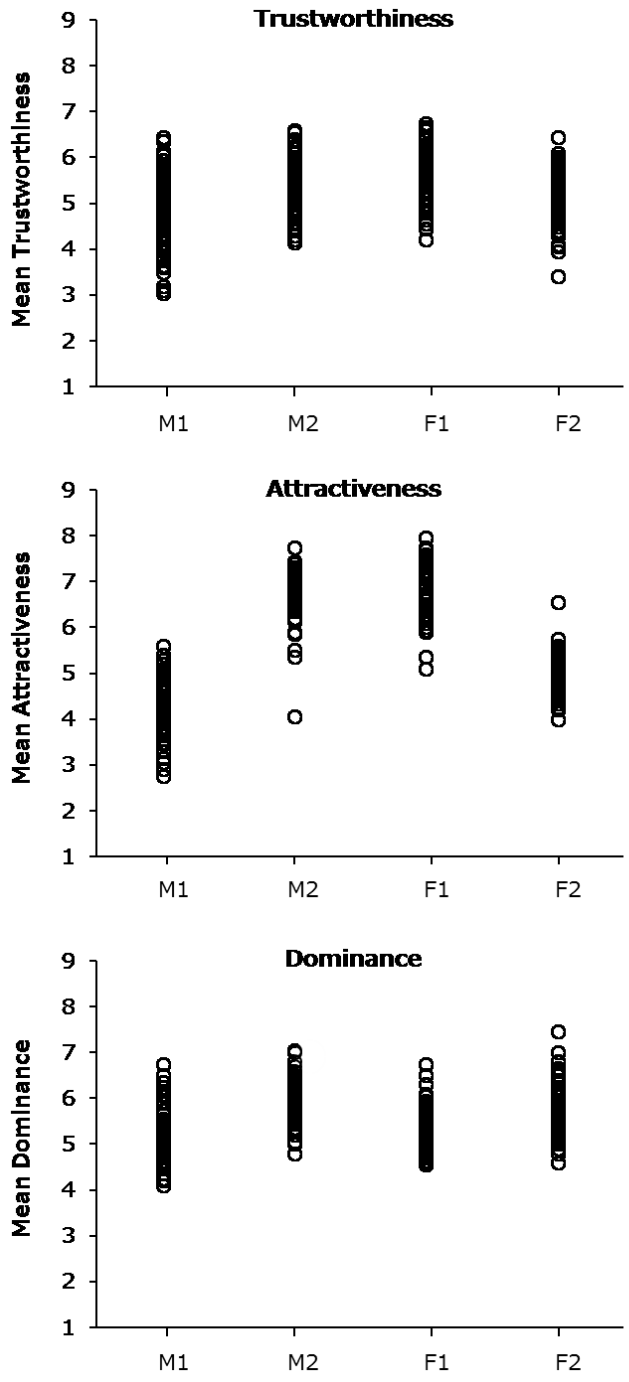

Figure 5: *Mean ratings of 100 images for each of four people (males M1 and M2, females F1 and F2) added to the set of stimuli used in Study 2. Ratings are shown for trustworthiness (top), attractiveness (middle) and dominance (bottom) for each identity. Each data point represents the average rating for each image. As for Study 1, there are again substantial differences in the ratings given to different images of the same person for all three judgements.*

Using image properties to predict impressions

Our regression models were each based either on 80 images of the same face, which we will call within-identity models, or on 80 images that included 16 face identities, which we call cross-identity models. These models were then tested for their ability to predict the variability in ratings of trustworthiness, attractiveness or dominance across sets of randomly selected images from the training sets (i.e. images of a single identity for the within-identity model and images of different identities for the cross-identity model) as well as 20 novel images of one person's face. When testing the performance of the within-identity models with untrained images, these test sets were always new images of the regression-trained face. The identical test sets were then used to assess the generalisability of the cross-identity models created with images of multiple other faces. In order to provide an estimation of chance performance, we used the same training and test image sets for both the within- and cross-identity models and shuffled all social trait ratings.

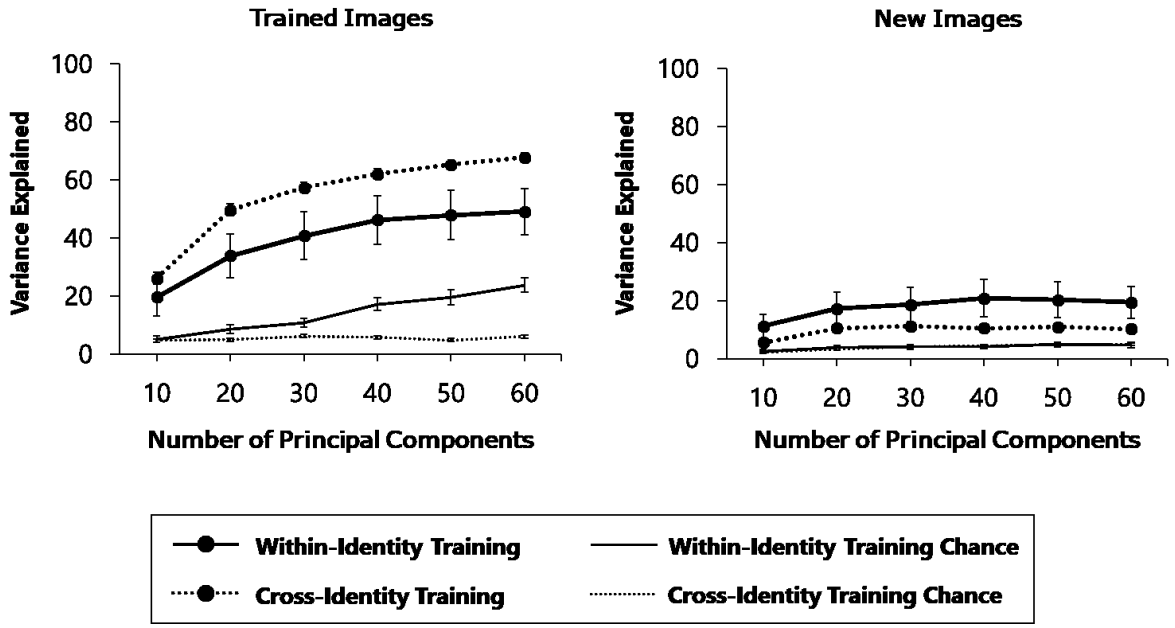

Figure 6: *Percentages of variance in participants' impressions in Study 2 that can be explained using different numbers of PCs for sets of trained images used to create each regression model (left panel) and in the cross-validation generalisation tests involving*

untrained novel images (right panel). Performance with trained images was tested by predicting social trait ratings from a randomly selected set of training images – these were images of the same person for the within-identity model and images of different identities for the cross-identity model. Performance with new images was tested by predicting social trait ratings from a set of images depicting the same identity (from the within-identity set) which were never included in the training set for either models. Mean performance across separate iterations involving the three modelled traits (trustworthiness, attractiveness and dominance) is shown, and error bars represent standard error. Performance with novel test images shows relatively modest benefit in both cases, but is consistently higher in the within-identity than the cross-identity condition.

In essence, then, we can directly compare performance in terms of the percentages of variance in impressions across within-identity and cross-identity models tested on exactly the same sets of novel images. Data are presented in Figure 6.

As in Study 1, we used up to 30 shape and 30 texture components. Shape components explained 99.7% and texture components explained 91.8% of the total variance on average across the four identities. For model creation, we used a combination of shape and surface PCs, as this was clearly optimal in Study 1. However, we do not show shape and surface PCs separately in Figure 6 because the overall level of generalisation test performance was substantially lower than for Study 1. Likewise, performance is averaged across the three different traits to show the key datum of the overall percentage of variance that can be explained by each type of model.

Both the within- and the cross-identity model performed significantly above chance. This was estimated with a Wilcoxon signed-ranks test based on data with 60 PCs. The results showed that the within-identity model performed significantly better than chance for both trained (Z = 5.14, $p < .001$) and new images ($Z = 5.51$, $p < .001$). The same was also true for the cross-identity model (trained images – $Z = 6.74$, $p < .001$; new images – $Z = 3.20$, $p = .001$). In fact, the same pattern holds even when using substantially fewer PCs. Supplementary Table 2 shows full statistics for each level of ‘N PCs’.

We will draw particular attention to two contrasting features of the data presented in Figure 6. First, better performance with the trained images could be achieved when multiple identities were used (the cross-identity modelled images in Figure 6, accounting for 68% of the

variance with 60 PCs) than when the models were trained on a single face (the within-identity modelled images, accounting for 49% of the variance with 60 PCs). Second, this pattern was reversed in the generalisation tests, with only 10% of the variance explained by cross-identity models and 19% by within-identity models.

The substantial drop in overall performance in the generalisation tests is in part likely to reflect the smaller sizes of the training and test image sets in Study 2, as compared to Study 1. More importantly, however, generalisation performance was higher for the models created from different images of the same identity (the within-identity models), with around 9% of additional variance explained. We tested this statistically using a Wilcoxon signed-ranks test to show better generalisation test performance from within-identity than cross-identity models based on 60 PCs ($Z = 3.25$, $p = .001$).

It thus seems that the generalisability of impressions may be slightly more reliable for a face that is well-known. We note though that the standard errors are substantially larger in the within-identity than the cross-identity condition for both the trained images and the generalisation test. This reflects differences between how well the impressions of different images of the four faces used in the within-identity condition could be modelled; some faces were easier to model than others.

Nonetheless, even the cross-identity models could account for some of the variance in impressions for the novel face identities used in the generalisation test; performance did not collapse to zero. A one-sample Wilcoxon signed-ranks test showed that the variance explained by the cross-identity model was significantly different from zero ($Z = 6.74$, $p < .001$). Hence, even though there may be some benefit to knowing a face well, at least some of the cues that create trait impressions are sufficiently consistent across different faces that personal knowledge of a face is not essential to forming impressions that will be consistent with those of other observers.

Study 3

Study 2 showed that learning to make trait attributions from a single face is to some extent identity-specific. Training based on images of a single face was able to capture less of the variance in impressions in the trained image set than was training based on images of a number of faces, yet generalisation to new images of the same face was slightly better than

generalisation to a different face. To that extent, this test at the limits of exposure to different identities (contrasting training based on a single face with training from a number of different faces) confirms that there does seem to be something idiosyncratic about impressions of a specific face. However, the circumstances in which one might learn to form impressions based entirely on a single individual do not replicate those prevailing in the natural environment, which is characterised by interactions with different people.

We therefore sought in Study 3 to assess the impact of learning to make trait inferences from a relatively small number of frequently encountered faces, as would be typical in the environment of most infants. To achieve this, we trained models on 80 images of each of 4 faces and tested generalisation to 20 new images of each of the same 4 faces (within-identity condition). We contrasted the performance of these 'familiar face impression' models to that of models trained with an equivalent total number of images (320 altogether) representing 16 different identities that did not appear in the novel test set (cross-identity condition). As an additional point of comparison, we added a 'mixed-identity' condition because the within-identity training set differed from the cross-identity training set both in the number of identities and in the number of images per identity. Like the cross-identity condition, the mixed-identity training set used 20 images of 16 faces, but (unlike the cross-identity condition) the training sets included some images of the faces that would appear in the novel test set. In this way, we could determine whether it is the presence of to-be-tested identities in the training set (as in the within-identity and mixed-identity conditions, but not in the cross-identity condition) that is a key factor underlying generalisation performance, or the number of trained images of the to-be-tested identities (which was larger in the within-identity condition than the mixed-identity condition, and zero in the cross-identity condition).

In each condition, then, the novel test images were identical; only the type of training (within-identity, cross-identity, or mixed-identity models) was manipulated.

Method

Stimuli, Participants, and Image rating task

Stimuli and ratings were all taken from those used in Study 1 and Study 2. By using 80 images of each of the 4 Study 2 set faces (320 images in total) and holding back 20 images of each of the Study 2 faces (80 in total) as novel test items we created stimuli for the within-

identity condition in which the model training and generalisation test items were of the same identities. This procedure was repeated 5 times, using different samples of 20 images from the 100 available for each face.

Each model was first tested with a set of randomly selected images which were included in the training set. These were images of the same identity for the within-identity model, images of different identities (different from the ones in the within-identity set) for the cross-identity model and images of different identities (including those in the within-identity set) for the mixed-identity model. The critical test, however, was how well these models could generalise to completely novel images. The same sets of novel test images were used for the cross-identity and mixed-identity conditions, but with regression models based on different training regimes. For the cross-identity condition, we used a model trained on the 320 images from Study 1 (20 images of 16 faces) that were of faces not included in the 80-item generalisation test sets. In the mixed-identity condition, we used a model trained on 20 images of each of the 4 faces that would appear in the appropriate generalisation test set together with 20 images of 12 other faces that were not in the generalisation test set.

As for Study 1, in order to ensure a fair comparison of levels of performance between trained images and novel test images across each training regime, we measured each model's performance for the 320 trained images using a random sample of 80 of these, thus equating the number of images used to measure performance in each case (always 80 images from the training set and 80 untrained novel test images).

Image PCA and Regression Models

Image PCA and the creation of regression models was carried out in the same way as for Study 1, involving model training on the sets of 320 images, randomly sampling 80 of these modelled images to measure training performance, and cross-validation tests for generalisation with the sets of 80 untrained novel images.

Results and Discussion

Our principal interest was in generalisation from within-identity models trained on the same small set of faces as the novel test images (4 identities in total), from mixed-identity models trained on a set of faces (16 identities in total) that included exemplars from the 4 faces used

to create the novel test images, or from cross-identity models trained on a larger set of different faces (16 identities in total) from the novel test images. As in Studies 1 and 2, we used up to 30 shape and texture components. Performing a PCA on images of all identities together revealed that the first 30 shape and texture components explained 99.5% and 88.2% of the overall variance respectively.

Using image properties to predict impressions

Our regression models were each based either on 80 images of each of 4 faces, which we call within-identity models, or on 20 images that included 16 face identities, which we call mixed-identity or cross-identity models. These models were then tested for their ability to predict the variability in trait ratings of trustworthiness, attractiveness or dominance across sets of trained images (these were images of the same identity for the within-identity model and images of different identities for the mixed- and cross-identity models) and untrained novel test images. The novel generalisation test images were the same across the three types of training condition, but in the within-identity condition these novel test images were of faces that had exclusively formed the training set, in the mixed-identity condition they were of faces that had formed part of the training set, and in the cross-identity condition they were of faces that had not been in the training set. In order to provide an estimation of chance performance we used the same training and test image sets for the within-, cross- and mixed-identity models and shuffled all social trait ratings.

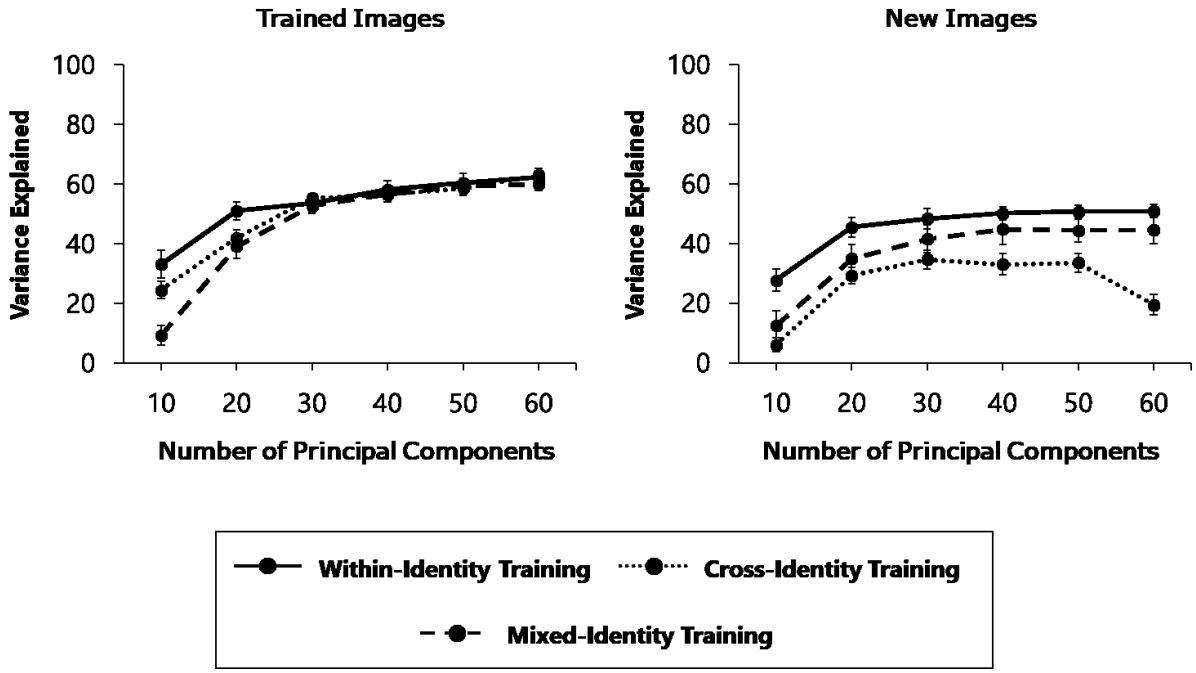

Figure 7: *Percentages of variance in participants' impressions in Study 3 that can be explained using different numbers of PCs for sets of trained images used to create each regression model (left panel) and in the cross-validation generalisation tests involving novel images of 4 faces (right panel). The within-identity model was trained on images of 4 different identities, the cross-identity model was trained on images of 16 other identities and the mixed model was trained on images of 16 identities which included the 4 identities from the within-identity set. Performance with trained images was tested by predicting social trait ratings from a randomly selected set of training images – these were images of the same person for the within-identity model and images of different identities for the cross- and mixed-identity models. Performance with new images was tested by predicting social trait ratings from a set of images depicting the same identity (from the within-identity set) which were never included in the training set for any of the models. Mean performance across separate iterations involving the three modelled traits (trustworthiness, attractiveness and dominance) is shown, and error bars represent standard error.*

Results are presented in Figure 7. As for Study 2, the data are averaged across the three trait dimensions to focus on the key points. They are also shown only for the combined analysis based on shape and surface texture PCs, but we note that separate analyses simply concurred with what we found in Study 1; namely that shape PCs derived from fiducial locations were

more effective than surface texture PCs, but a combination of both (as presented in Figure 7) achieved the best results.

Our estimation of chance showed highly similar values across the three models as well as across all levels of PC numbers. The mean chance performance for both trained and new generalisation images was 1% with maximum values of only 2%. Wilcoxon signed-ranks tests showed that all models performed significantly above chance levels both for trained and new generalisation images (all Zs = 3.41, all ps = .001).

As can be seen, the trained models were able to account for substantial and closely comparable proportions of the variance in trait impressions in the training sets; 62% overall for within-identity training using 60 PCs, 60% overall for mixed-identity training, and 63% overall for cross-identity training. However, there was a clear difference in the generalisation tests using novel images, with 51% of the variance in novel test images explicable following within-identity training using 60 PCs, 45% overall following mixed-identity training, and 19% overall following cross-identity training. Wilcoxon matched-pairs signed-ranks tests based on generalisation test data from 60 PCs showed that a significantly higher amount of variance was explained by within-identity training compared to cross-identity training (Z = 3.18, p = .003) and by mixed-identity training compared to cross-identity training (Z = 3.01, p = .006), whilst within-identity and mixed-identity models did not differ significantly (Z = 1.02, p > .05).

Moreover, there were clear signs of overfitting in the cross-identity training condition, with performance to novel test images actually declining across higher numbers of PCs. That said, it remains the case that on average, 33% of the variance in trait impressions of entirely novel faces could be successfully modelled from between 20 and 50 shape and texture PCs in the cross-identity condition. Moreover, Wilcoxon matched-pairs signed-ranks tests based on generalisation test data from 50 PCs showed the same pattern as found for 60 PCs, with a significantly higher proportion of variance explained by within-identity training compared to cross-identity training (Z = 2.27, p = .046) and by mixed-identity training compared to cross-identity training (Z = 2.73, p = .018), whilst within-identity and mixed-identity models did not differ significantly (Z = 1.02, p > .05). The Holm-Bonferroni correction (1979) was applied to account for the multiple comparisons. Supplementary Table 3 gives statistical comparisons between conditions at each level of ‘N PCs’ (10 to 60).

General Discussion

A paradox in recent studies of face perception has been that while people can readily form reasonably consistent (consensual) trait impressions from images of unfamiliar faces, their perception of unfamiliar face identity is both error-prone and inconsistent between different observers (Jenkins et al., 2011; Young, 2018; Young and Burton, 2017, 2018a, 2018b). Our aim was to understand this difference between relatively consensual trait impressions and the inconsistent perception of the identities of unfamiliar faces.

Our work is grounded in the observation that ambient images of faces can be highly variable, with substantial differences resulting from changes in lighting, pose, expression and so on. Analyses of this variability have shown that it can be substantial, even for images of the same face, but that it is also to some extent identity-specific in the sense that the way in which one person's face varies may differ from how someone else's face varies (Burton et al., 2016). This variability can therefore create problems in recognising the identities of unfamiliar faces, whose dimensions of variability are by definition unknown to the observer. Learning to recognise a face thus involves learning to cope with and make use of its idiosyncratic variability (Burton et al., 2016; Kramer, Young et al., 2017; Kramer et al., 2018).

This image variability also offers a substantial challenge to creating models that can simulate the impressions of human observers, since different images of the same face can create very different subjective impressions (Jenkins et al., 2011; Sutherland, Young et al., 2017; Todorov & Porter, 2014), as was evident in the ratings collected here (Figure 2 and Figure 5). Moreover, while theoretical models of facial impression formation converge on the idea that most traits fall along two or three underlying dimensions, they tend to use different descriptor labels for the dimensions themselves. The seminal study by Oosterhof and Todorov (2008) called the two dimensions they identified valence/trustworthiness and dominance, whereas Sutherland et al. (2013) called their three dimensions approachability, youthful-attractiveness and dominance. However, all researchers accept that applying verbal labels to the dimensions revealed by PCA or factor analysis will always be somewhat imprecise, and trustworthiness and approachability are in fact close correlates of each other (Sutherland et al., 2013), so there is no substantive disagreement. On this basis, we chose to use ratings of trustworthiness, attractiveness and dominance as proxies for each of the main dimensions.

In Study 1, we were able to create PCA-based regression models that could on average

account for 51% of the variance in consensual impressions of entirely novel, naturally-occurring, images across these dimensions. This is comparable to previous work by Vernon, Sutherland, Young and Hartley (2014), who developed an approach that could capture 58% of the variance in human observers' impressions of novel ambient images using a linear neural network trained on an arbitrary set of 65 different physical attributes measured in each image, such as eye width, eyebrow width, mouth width and eyes-to-mouth distance. However, while Vernon et al.'s (2014) approach showed that it is possible to model subjective impressions of highly varied everyday images from objectively specified attributes, it was subject to the limitation that the choices of these 65 attributes represented arbitrary decisions of the experimenters. These choices may have failed to represent all of the information potentially present in the images and they also imposed constraints that meant that the model was not directly image-based as it involved an intermediate step of calculating the 65 attributes.

From PCA-based models, Study 1 was able to achieve a level of generalisation performance approaching that reported by Vernon et al. (2014) even though our training sets of 320 images were smaller than the sets of 800 images Vernon et al. used to train their neural networks. However, the most important difference between the procedures used in Study 1 and by Vernon et al. (2014) is not the number of images used to create the models, but rather the absence in Study 1 of any intervening image analysis beyond a statistical description of the variability of the images themselves. PCA makes no assumptions concerning which features or attributes might be important, or indeed whether the critical information can be described linguistically (e.g. in terms of attributes or features) at all.

That a PCA-based approach can work so well shows that much of the information used to create consensual trait impressions can be found in the images themselves. Theoretical approaches to facial impression formation often emphasise inferences based on stereotypes involving gender or age (Macrae & Bodenhausen, 2000; Oldmeadow et al., 2013; Quinn & Macrae, 2011; Sutherland, Young, Mootz & Oldmeadow, 2015) and the importance of facial expression (Montepare & Dobish, 2003, 2014; Oosterhof & Todorov, 2008; Said, Sebe, & Todorov, 2009; Sprengelmeyer et al., 2016; Zebrowitz, Kikuchi, & Fellous, 2007). While such higher-order inferential factors undoubtedly play a role in impression formation, our data show that much of the variance in impressions can be modelled without needing to use them as explicit mediators. This may in part be due to the fact that PCA analyses dimensions

of physical variability because, as already noted, it is clear that it is covariation between cues that determines facial impressions rather than individual cues *per se* (Santos & Young, 2011; Todorov, 2017; Young, 2018). For example, although smiling is often regarded as a cue for approachability, smiling can also make a person look attractive and even in some circumstances dominant. It is the type of smile and especially the way that it is combined with other cues such as face shape, age, skin colour and head orientation that creates the overall impression. The combination of regression and PCA may therefore be a valuable way of finding such covariation. In understanding the impact of cue covariation, however, it is useful to note that the techniques used here and by Vernon et al. (2014) are mainly linear; they do not presuppose (and would not find) more complex non-linear interactions. Their relative success suggests that linear models may be sufficient to mimic much of what our brains can do. In fact, Todorov and Oosterhof (2011) directly compared the performance of linear and nonlinear (quadratic) models and showed only limited improvement in the amount of variance explained by a quadratic model. In this respect, it is interesting that fMRI studies have shown that face-responsive brain regions also track relatively linear changes in face properties (Baseler, Young, Jenkins, Burton & Andrews, 2016). In making this point, we are not however seeking to claim that the human brain uses PCA as such - the point is only that PCA offers a useful way of demonstrating the presence of information that the brain can exploit.

That said, there remains a substantial proportion of variance that has not been captured by these linear approaches. How much of this unexplained variance is meaningful and how much is simply measurement error is unknown, because there is no objective criterion for whether an impression is 'correct'; what we are testing is agreement between a regression model and the average impressions of an independent group of observers. In this respect, it is important to emphasise that whilst understanding consensual impressions is important, these are not everything that needs to be explained. Although many studies report substantial consistency of facial impressions across different observers (as we do here), forming a consensual core of 'shared taste' that is consistent across most observers, it is also acknowledged that there is a significant contribution from individual differences that correspond to 'private taste' (Germine et al., 2015; Hehman, Sutherland, Flake, & Slepian, 2017; Hönekopp, 2006; Kramer et al., 2018). This is evident in the data we report here, where inter-rater agreement about the overall group values, as measured with Cronbach's alpha was high for all three social attributes. In contrast, Kendall's coefficient of concordance across

different observers, W, was noticeably lower, even though it did indicate an underlying core of significant agreement. Averaging the impressions of each image across multiple observers (as we did here) will reduce the impact of these observer differences, but not fully eliminate them.

It is also important to note that very similar impressions can be based on purely auditory cues, with research showing that the relative importance of facial and vocal information varies as a function of the specific social trait in question (Mileva, Tompkinson, Watt, & Burton, 2018; Rezlescu et al., 2015). Understanding how auditory and visual information are combined to create an almost instantaneous impression of a *person* represents an important theoretical task (cf. Young, 2018).

Turning back to the visual information itself, although shape PCs used in isolation were more effective than texture PCs in modelling facial impressions, a combination of shape and surface texture PCs achieved the best results. However, it is worth emphasising that our approach involves analysing the PCs of the shapes and textures of the images themselves, which are not simply the same as the shapes or textures of the faces - the same face can have different fiducial positions and different surface texture patterns in different images. Moreover, although widely used in the computer science and psychology literatures (Burton et al., 2001; Sutherland, Rhodes et al., 2017), this procedure does not create a perfect separation between shape and surface texture information. For example, shape from shading cues do not typically influence 2D fiducial locations and hence become treated as a surface property of each image (Sormaz, Watson, Smith, Young & Andrews, 2016; Sormaz, Young & Andrews, 2016; Kramer et al., 2018).

Nonetheless, the usefulness of image shape PCs based on fiducial locations for modelling facial impressions stands in marked contrast to their lack of importance to classifying familiar identity. In a modelling study using the same type of ambient images, Kramer et al. (2018) showed that a combination of PCA and Linear Discriminant Analysis (LDA) could be used to recognise a number of trained face identities in a set of highly variable ambient images, and that 80 training images (equivalent to the number of images we used to create the regression models of impressions in Study 2) were sufficient to reach asymptotic levels of face identity recognition (close to 100% correct) across nearly 4,000 images containing many different identities (Kramer et al., 2018). However, Kramer et al. (2018) found that excellent

performance for recognising identity could only be achieved with a model based on surface texture PCs; the performance of image shape PCs for recognising identity was poor (around 3% correct overall). This poor performance in familiar face recognition from shape PCs almost certainly reflects the fact that the 2D image fiducial locations themselves are altered by changes in viewpoint, pose and expression, rendering them too unstable to find identity-specific properties (Burton, Schweinberger, Jenkins & Kaufmann, 2015; Kramer et al., 2018). However, the factors that create this instability at the fiducial level (changes in viewpoint, pose and expression) are of course highly relevant to facial impressions (Jenkins et al., 2011; Sutherland, Young et al., 2017), so it makes sense that image shape PCs are useful to modelling these impressions. There is a profound difference between the demands of face identity recognition (for which the impact of within-identity image differences needs to be minimised) and the creation of trait impressions (for which within-identity image differences carry a great deal of meaningful information) (Young, 2018).

Our contention is therefore that the roles of within-face and between-face variability are of critical theoretical importance. The reason why Kramer et al. (2018) used a combination of PCA and LDA for face recognition lay in Burton et al.'s (2016) finding that the statistical variability of a face across different images is to some extent identity-specific; that is, the ways in which Vladimir Putin's face varies across different images will be different from the ways in which Donald Trump's face varies. Under these circumstances LDA offers a useful way to cluster together images of the same face identity; reshaping the underlying image-based PCA space into a representational space organised around different face identities that can bring images of Putin (or of Trump) closer together. However, this PCA+LDA technique mainly serves to recognise faces from a trained set of 'familiar' identities. Like human observers, the PCA+LDA model performs relatively poorly in recognising unfamiliar face identities (Kramer et al., 2018).

In Study 2 and Study 3, we also found some evidence of identity-specific variability in first impressions. In Study 2, training involving a single face identity led to better generalisation of performance to novel images of that face than did training from different identities even though training performance was better for multiple identities (the cross-identity training condition in Study 2) than for a single face (within-identity training in Study 2). However, although above-chance in Study 2, ability to successfully predict impressions of novel images of a specific face was relatively limited; accounting for 19% (following within-identity

training) or 10% (following cross-identity training) of the variance in impressions. A simple analogy that might be used to think about this phenomenon would be that if you have a friend who does not smile very often, then one of their smiles will carry greater weight than that of another friend who smiles most of the time. Interpretation of the meaning of a smile will then be, in part, identity-specific.

Pushing this analogy further, however, while interpretation of the meaning of a smile will be in part identity-specific, smiling itself remains something that is interpretable across different individual faces. Indeed, in Study 3 we noted overall levels of generalisation performance comparable to those obtained in Study 1 from training involving a relatively small set of face identities. Using the same analogy, any identity-specific variability in the probability of smiling will not be such as to entirely prevent generalisation of the implications of a smile to other identities. In other words, although there is some idiosyncratic variation, much the same combinations of cues may well signal trustworthiness, attractiveness or dominance in any face, be it familiar (in our case, a face on which the regression model was trained) or unfamiliar (i.e. an image of a novel face). This point is also clear from Vernon et al.'s (2014) study, where an overall level of performance comparable to our Study 1 and Study 3 was achieved with a model based entirely on a single ambient image of each face, i.e. without representing idiosyncratic face variability at all. Nonetheless, our findings here suggest that it is likely useful to be exposed to at least a small number of different faces to arrive easily at relatively stable interpretations of the meaning of different cue combinations.

Our findings therefore help resolve the paradox of why people can so readily form consensual impressions of unfamiliar faces when their perception of unfamiliar face identity is both error-prone and inconsistent between different observers (Young, 2018; Young & Burton, 2017, 2018a, 2018b). Although we have shown (in Study 2 and Study 3) that, like identity, the variability that underlies facial trait impressions is to some extent idiosyncratic, relatively speaking it is less person-specific than the variability that underlies face identity; there is useful information that can be generalised to any face image in order to interpret its trustworthiness, attractiveness and dominance to some extent.

These insights contrast with a relatively traditional approach to face recognition, which has often sought to use highly standardised images in an attempt to minimise the impact of ‘nuisance’ variation from factors such as lighting, camera differences, etc. However, we have

shown elsewhere that this approach can obscure our understanding (Burton, 2013; Young, 2018). By using the full range of ambient images of the type people recognise every day in their newspapers, televisions, and online, we can preserve the natural within-person variations that characterise our real world experience of faces. Far from being a nuisance, this variability is a necessity in allowing us to find consistent cues for recognising familiar face identity (Bruce, 1994; Burton, 2013) and as we show here it is critical to other aspects of face perception as well (Bruce & Young, 2012; Sutherland et al., 2013; Vernon et al., 2014).

From a broader standpoint, facial impressions seem to fall along key dimensions that other primate species also use to evaluate conspecifics (Fiske, Cuddy & Glick, 2007; Oosterhof & Todorov, 2008; Sutherland, Oldmeadow & Young, 2016). Trustworthiness or approachability can be considered as involving the appraisal of intentions to help or harm, attractiveness is linked to mechanisms of sexual selection, and dominance involves an appraisal of ability to carry out intentions. What is different for humans is that we are exposed to such large numbers of unfamiliar individuals and, as Todorov (2017) emphasises, this is in evolutionary terms a relatively recent cultural phenomenon. An interesting idea is therefore that mechanisms underlying first impressions are initially established in relatively small groups of familiar people encountered in infancy and childhood (cf. Pascalis et al., 2014; Lee, Quinn & Pascalis, 2017) where they may well have relatively high validity (if your mum looks unapproachable, it's not the best time to ask for more pocket money). They are then overgeneralised to strangers' faces and mistakenly considered to represent stable traits in a similar manner to the fundamental attribution error of interpreting others' behaviour as reflecting enduring personality characteristics (cf. Todorov, 2017). Our findings help to elucidate the perceptual underpinnings of this process.

Supplementary Materials

All supplementary data (raw ratings for both the between and the within person image sets and the accompanying shape and texture PCA projections for each image) and all supplementary tables referred to in the text can be obtained from the Open Science Framework (doi: 10.17605/OSF.IO/UQP9B) at:

https://osf.io/uqp9b/?view_only=adb8be2e27404bedb7a419b7395c8d41

References

Baseler, H. A, Young, A. W., Jenkins, R., Burton, A. M., & Andrews, T. J. (2016). Face-selective regions show invariance to linear, but not non-linear, changes in the spatial configuration of faces. *Neuropsychologia, 93*, 76-84. https://doi.org/10.1016/j.neuropsychologia.2016.10.004

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. doi: 10.1163/156856897X00357

Bruce, V. (1994). Stability from variation: The case of face recognition. *Quarterly Journal of Experimental Psychology, 47A*, 5-28. https://doi.org/10.1080/14640749408401141

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3)*,* 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruce, V., & Young, A. (2012). *Face perception*. Hove, East Sussex: Psychology Press.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology, 66*(8), 1467-1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256-284. https://doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*(1), 202–223. https://doi.org/10.1111/cogs.12231

Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image formats. *Vision Research, 41*(24), 3185–3195. https://doi.org/10.1016/S0042-6989(01)00186-9

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments against a configural processing account of familiar face recognition. *Perspectives on Psychological Science, 10*(4), 482-496. https://doi.org/10.1177/1745691615583129

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286-291. https://doi.org/10.3758/BRM.42.1.286

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research, 41*(9), 1179-1208. https://doi.org/10.1016/S0042-6989(01)00002-5

Cohn, J. F., Schmidt, K., Gross, R., & Ekman, P. (2002). Individual differences in facial

expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces* (pp. 491–496). Los Alamitos, CA: IEEE Computer Society.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Craw, I. (1995). A manifold model of face and object recognition. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition* (pp.183-203). London: Routledge.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A. M., Smoller, J. W., Kwok, H., … Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology, 25*(20), 2684–2689. https://doi.org/10.1016/j.cub.2015.08.048

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 330-337. https://doi.org/10.1016/S1364-6613(00)01519-9

Hehman, E. A., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology, 113*(4), 513-529. http://dx.doi.org/10.1037/pspa0000090

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 199–209. doi: 10.1037/0096-1523.32.2.199

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Kaufmann, J. M., & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception, 33*(4), 399-408. https://doi.org/10.1068/p5083

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36*,

ECVP Abstract Supplement.

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods, 49*(6), 2002-2011. https://doi.org/10.3758/s13428-016-0837-7

Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PloS One, 13*(8), e0202655. https://doi.org/10.1371/journal.pone.0202655

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition, 172*, 46-58. https://doi.org/10.1016/j.cognition.2017.12.005

Kramer, R. S. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review*, *124*(2), 115-129. http://dx.doi.org/10.1037/rev0000048

Lee, K., Quinn, P. C., & Pascalis, O. (2017). Face race processing and racial bias in early development: a perceptual-social linkage. *Current Directions in Psychological Science, 26*(3), 256-262. https://doi.org/10.1177/0963721417690276

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120. https://doi.org/10.1146/annurev.psych.51.1.93

Mileva, M., & Burton, A. M. (2018). Smiles in face matching: Idiosyncratic information revealed through a smile improves unfamiliar face matching performance. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12318

Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2018). Audiovisual integration in social evaluation. *Journal of Experimental Psychology: Human Perception and Performance, 44*(1), 128-138. http://dx.doi.org/10.1037/xhp0000439

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27*(4), 237-254. https://doi.org/10.1023/A:1027332800296

Montepare, J. M., & Dobish, H. (2014). Younger and older adults' beliefs about the experience and expression of emotions across the life span. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 69*(6), 892-896. https://doi.org/10.1093/geronb/gbt073

Oldmeadow, J. A., Sutherland, C. A. M., & Young, A. W. (2013). Facial stereotype visualization through image averaging. *Social Psychological and Personality Science, 4*(5), 615-623. https://doi.org/10.1177/1948550612469820

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–70. https://doi.org/10.1016/j.tics.2014.09.007

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, USA, 105*(32)*,* 11087-11092. https://doi.org/10.1073/pnas.0805664105

Pascalis, O., Loevenbruck, H., Quinn, P. C., Kandel, S., Tanaka, J. W., & Lee, K. (2014). On the links among face processing, language processing, and narrowing during development. *Child Development Perspectives*, *8*(2), 65–70. https://doi.org/10.1111/cdep.12064

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. doi: 10.1163/156856897X00366

Quinn, K. A., & Macrae, C. N. (2011). The face and person perception: Insights from social cognition. *British Journal of Psychology, 102*(4), 849–867. https://doi.org/10.1111/j.2044-8295.2011.02030.x

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant voices and attractive faces: The contribution of visual and auditory information to integrated person impressions. *Journal of Nonverbal Behavior, 39*(4), 355–370. http://dx.doi.org/10.1007/s10919-015-0214-8

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*, 199–226. https://doi.org/10.1146/annurev.psych.57.102904.190208

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*(2), 260-264. http://dx.doi.org/10.1037/a0014681

Santos, I. M., & Young, A. W. (2011). Inferring social attributes from different face regions: Evidence for holistic processing. *Quarterly Journal of Experimental Psychology, 64*(4), 751-766. https://doi.org/10.1080/17470218.2010.519779

Sormaz, M., Watson, D. M., Smith, W. A. P., Young, A. W., & Andrews, T. J. (2016a). Modelling the perceptual similarity of facial expressions from image statistics and neural responses. *NeuroImage, 129*, 64-71. https://doi.org/10.1016/j.neuroimage.2016.01.041

Sormaz, M., Young, A. W., & Andrews, T. J. (2016b). Contributions of feature shapes and surface cues to the perception of facial expressions. *Vision Research, 127*, 1-10.

https://doi.org/10.1016/j.visres.2016.07.002

South Palomares, J. K., & Young, A. W. (2018). Facial first impressions of partner preference traits: Trustworthiness, status, and attractiveness. *Social Psychological and Personality Science,* doi:10.1177/1948550617732388.

South Palomares, J. K., Sutherland, C. A. M., & Young, A. W. (2018). Facial first impressions and partner preference models: Comparable or distinct underlying structures? *British Journal of Psychology, 109*(3), 538-563. https://doi.org/10.1111/bjop.12286

Sprengelmeyer, R., Young, A. W, Baldas, E-M., Ratheiser, I., Sutherland, C. A. M., Müller, H-P., et al. (2016). The neuropsychology of first impressions: Evidence from Huntington's disease. *Cortex, 85,* 100-115. https://doi.org/10.1016/j.cortex.2016.10.006

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105-118. https://doi.org/10.1016/j.cognition.2012.12.001

Sutherland, C. A. M., Oldmeadow, J. A., & Young, A. W. (2016). Integrating social and facial models of person perception: Converging and diverging dimensions. *Cognition, 157*, 257-267. https://doi.org/10.1016/j.cognition.2016.09.006

Sutherland, C. A. M., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *Social Psychological and Personality Science*, *8*(5), 538-551. https://doi.org/10.1177/1948550617697176

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology, 106*(2), 186-208. https://doi.org/10.1111/bjop.12085

Sutherland, C. A. M., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgments are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*(2), 397-415. https://doi.org/10.1111/bjop.12206

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton, NJ: Princeton University Press.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626. doi: 10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions

from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces [social sciences]. *IEEE Signal Processing Magazine, 28*(2), 117-122. doi: 10.1109/MSP.2010.940006

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*(7), 1404–1417. https://doi.org/10.1177/0956797614532474

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences, USA, 111*, E3353-E3361. https://doi.org/10.1073/pnas.1409860111

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision, 9*(11), 1-13. doi: 10.1167/9.11.12

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Young, A. W. (2018). Faces, people and the brain: The 45th Sir Frederic Bartlett Lecture. Q*uarterly Journal of Experimental Psychology, 71*(3), 569-594. https://doi.org/10.1177/1747021817740275

Young, A. W., & Burton, A. M. (2017). Recognizing faces. *Current Directions in Psychological Science, 26*(3), 212-217. https://doi.org/10.1177/0963721416688114

Young, A. W., & Burton, A. M. (2018a). Are we face experts? *Trends in Cognitive Sciences, 22*(2), 100-110. https://doi.org/10.1016/j.tics.2017.11.007

Young, A. W., & Burton, A. M. (2018b). What we see in unfamiliar faces: A response to Rossion. *Trends in Cognitive Sciences, 22*(6), 472-473. https://doi.org/10.1016/j.tics.2018.03.008

Zebrowitz, L. A., Kikuchi, M., & Fellous, J. M. (2007). Are effects of emotion expression on trait impressions mediated by babyfaceness? Evidence from connectionist modeling. *Personality and Social Psychology Bulletin, 33*(5), 648-662. https://doi.org/10.1177/0146167206297399